

Document Version

Final published version

Licence

Dutch Copyright Act (Article 25fa)

Citation (APA)

Roque, E. B., & Cruz, L. (2025). Energy Aware Development of Neuromorphic Implantables: From Metrics to Action. In *Proceedings - 2025 11th International Conference on ICT for Sustainability, ICT4S 2025* (pp. 198-208). IEEE.
<https://doi.org/10.1109/ICT4S68164.2025.00028>

Important note

To cite this publication, please use the final published version (if applicable).
Please check the document version above.

Copyright

In case the licence states "Dutch Copyright Act (Article 25fa)", this publication was made available Green Open Access via the TU Delft Institutional Repository pursuant to Dutch Copyright Act (Article 25fa, the Taverne amendment). This provision does not affect copyright ownership.
Unless copyright is transferred by contract or statute, it remains with the copyright holder.

Sharing and reuse

Other than for strictly personal use, it is not permitted to download, forward or distribute the text or part of it, without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license such as Creative Commons.

Takedown policy

Please contact us and provide details if you believe this document breaches copyrights.
We will remove access to the work immediately and investigate your claim.

Energy Aware Development of Neuromorphic Implantables: From Metrics to Action

Enrique Barba Roque[✉]
Delft University of Technology
Delft, The Netherlands
enrique@ebarba.com

Luis Cruz[✉]
Delft University of Technology
Delft, The Netherlands
L.Cruz@tudelft.nl

Abstract—Spiking Neural Networks (SNNs) and neuromorphic computing present a promising alternative to traditional Artificial Neural Networks (ANNs) by significantly improving energy efficiency, particularly in edge and implantable devices. However, assessing the energy performance of SNN models remains a challenge due to the lack of standardized and actionable metrics and the difficulty of measuring energy consumption in experimental neuromorphic hardware. In this paper, we conduct a preliminary exploratory study of energy efficiency metrics proposed in the SNN benchmarking literature. We classify 13 commonly used metrics based on four key properties: Accessibility, Fidelity, Actionability, and Trend-Based analysis. Our findings indicate that while many existing metrics provide useful comparisons between architectures, they often lack practical insights for SNN developers. Notably, we identify a gap between accessible and high-fidelity metrics, limiting early-stage energy assessment. Additionally, we emphasize the lack of metrics that provide practitioners with actionable insights, making it difficult to guide energy-efficient SNN development. To address these challenges, we outline research directions for bridging accessibility and fidelity and finding new Actionable metrics for implantable neuromorphic devices, introducing more Trend-Based metrics, metrics that reflect changes in power requirements, battery-aware metrics, and improving energy-performance tradeoff assessments. The results from this paper pave the way for future research on enhancing energy metrics and their Actionability for SNNs.

Index Terms—neuromorphic computing, spiking neural networks, energy metrics, benchmarking, implantables, artificial intelligence

I. INTRODUCTION

Neuromorphic computing and Spiking Neural Networks (SNNs) are an emerging alternative to traditional Von Neumann computing¹ and Artificial Neural Networks that aim to improve energy efficiency while conserving high accuracy. This paradigm redesigns Neural Networks to more closely mimic the behaviour of the brain, encoding information across time in binary spikes, reducing the complexity of computation, and increasing parallelization. These SNNs work in conjunction with the specialized neuromorphic hardware to reduce energy usage.

Design and fabrication of neuromorphic hardware are still experimental and in early stages. SNN models can be built in Python using libraries built on top of PyTorch, and can be later deployed to neuromorphic hardware. While these models can

be run on traditional CPU and GPU hardware, this execution will not reflect the energy efficiency gains that can be obtained from the neuromorphic hardware. However, having access to neuromorphic hardware for deploying and testing the efficiency of the model is rather difficult, given the experimental nature of its components. Therefore, being able to benchmark and detect inefficiencies in SNNs before deploying them into hardware is a relevant problem in the community.

Different works have tried to define standard SNN benchmarks, and metrics to evaluate accuracy and energy efficiency of these models [1, 2, 3]. However, while some of these metrics might be useful for ranking and comparison among different SNN architectures, they are difficult to interpret by developers.

Hence, there is an important gap in helping ML practitioners use the right metrics that allow them to select the most energy-efficient models that fulfill the requirements of the project. We anticipate that there is no one-size-fits-all metric for energy efficiency, as it is highly dependent on the use case. Some use cases might require extending the lifetime of the battery as much as possible, while other use cases might focus on saving electricity bills, and others might prioritize reducing carbon emissions. Hence, different stakeholders have different needs, and different use cases require different perspectives.

This is particularly important in the context of this study, focused on the Smart Edge Lab for Healthcare (SELF lab) at TU Delft, where we are developing medical implants for early epilepsy detection². Such a device brings a multi-disciplinary team together that works at different layers of the implant: model development, hardware controller, hardware design, patient intervention, etc. Thus, we need different stakeholders with different disciplinary backgrounds to be able to communicate and make decisions based on energy metrics.

In this exploratory paper, we present an overview of the state-of-the-art metrics for energy efficiency proposed in SNN benchmarking literature. Given that each metric has its pros and cons, we classify them according to different perspectives: Accessibility, Fidelity, Actionability, and Trend-Based. We find that most metrics do not provide actionable insights to developers, and propose some research directions for alternative metrics that can improve energy assessment in the early stages of SNN model development.

¹Traditional computing architecture based on shared memory and sequential execution of instructions.

²Epilepsy is a neurological disease that manifests as a brain-wide phenomenon.

II. BACKGROUND

This Section introduces the necessary technical background to understand this paper. We explain the main differences between Artificial Neural Networks and Spiking Neural Networks, and how the latter are directly supported by a new paradigm of neuromorphic hardware design.

A. ANNs vs SNNs

Advancements in the field of artificial intelligence have been predominantly influenced by the application of Artificial Neural Networks (ANN). Essentially, ANN models are composed of layers of interconnected neurons where each neuron receives inputs – typically represented as floating point numbers – processes these inputs through multiply-accumulate operations (MAC operations) and sends them to subsequent neurons through weighted connections. Learning methods, normally based on gradient descent and backpropagation, are used to train weights and improve the accuracy of the model.

Although the idea is not new, most advancements in this technology come from the ability to increase model complexity with more processing power [4], adding more layers and more complex neurons, increasing the total arithmetic operations and, therefore, increasing energy consumption. This makes ANNs suboptimal for low-power scenarios, like wearable devices or edge computing.

To tackle this energy efficiency problem, Spiking Neural Networks have emerged as an alternative. Taking inspiration from the human brain, these neural networks encode information in the form of discrete binary spikes across time [5] rather than in the form of real numbers. One of the most popular neuron models is the Leaky Integrate and Fire (LIF), initially developed by Lapicque in 1907 [6] to explain neurons' behavior, and later implemented for SNNs. In this model, neurons keep an internal state – called membrane potential, referencing brain mechanisms – which increases when receiving spikes and makes the neuron fire a spike after a certain threshold. When combined with neuromorphic hardware, this translates to multiple advantages compared to traditional ANNs.

Mainly, this design makes SNN an event-driven model [7]. The activity in the hardware and energy usage happens mostly during spikes, which are relatively sparse. While a neuron is idle and not receiving or sending spikes, energy consumption is minimal.

Additionally, neuromorphic hardware is designed with high parallelization in mind. Processing and memory are co-located, forming one neuron, instead of having a central CPU and memory. Therefore, multiple neurons can be operating simultaneously, and, since the spikes are binary values, most operations are simple accumulations (AC operations), and the number of MAC operations, which are typically more expensive, is greatly reduced.

Currently, the development of SNN models takes place mainly in Python. There are multiple libraries built on top of the PyTorch framework [8], such as SNN-Torch [9] or SpikingJelly [10]. Combined with neuromorphic hardware, these SNNs can provide similar accuracy to ANN while using up to 100x less energy [11].

B. Neuromorphic hardware

To completely exploit the energy efficiency gains of SNNs, these need to be deployed in neuromorphic hardware. This kind of hardware physically implements the concept of spiking neurons to mimic neuron models like the previously mentioned LIF. In contrast to traditional Von Neumann architecture, this hardware usually includes collocated processing and memory. This leads to efficient event-driven parallel computing, where neurons can independently compute their results based on their inputs. Additionally, the architecture is easily scalable by chaining additional neurons or chips. Neuromorphic hardware can also be fully digital, using traditional binary computation, or mixed analog-digital design. For this kind of design, LIF neurons are implemented with a circuit that includes a capacitor that stores the neuron's potential, and a comparator that fires when this potential is over a threshold voltage.

An example of this is SpiNNaker [12], which is composed of 57,600 interconnected nodes, each with 18 of these processors and 128 MB of memory. Intel is also working on neuromorphic computing with Loihi and Loihi 2 [13].

BrainScale [14] is a hybrid analog implementation of this concept. Their paper also provides a comparison between some of the previously mentioned architectures and their own, as well as the efficiency of traditional ANN with GPU hardware. They show that neuromorphic computing can have gains of up to 101x compared to traditional ANNs.

In conclusion, it is hard to measure and compare the energy efficiency of SNNs when they are not deployed into neuromorphic hardware that aims to maximize this efficiency. For traditional ANNs, there are mature tools like RAPL [15] and MLPerf [16] to benchmark models. While these tools can be used for SNNs run in traditional hardware, the measurements obtained will hardly reflect the energy improvements that would be obtained with neuromorphic hardware.

III. RESEARCH CONTEXT

This paper stems from a real context in the SELF Lab where we need to deploy energy-efficient neural networks on implantable devices. Concretely, the overarching project aims to develop an implantable neural device that runs an AI model for the early detection of epilepsy events, and early neurostimulation to prevent the episode. A high-level illustration of the project is presented in Figure 1.

Therefore, the project brings together a multidisciplinary team with stakeholders who are involved across the entire device development pipeline. For example, part of the team works on hardware design and techniques to improve memristor hardware design [17], while another part is working on learning algorithms for SNNs [18]. Ultimately, healthcare professionals and epilepsy patients also play an important role as they want to minimize any risks involved with implanting and using the smart edge device.

In this context, energy consumption plays an important role: we want to maximize the lifetime of using the device without intervention. This means that the less energy the model spends, the less likely it is that the device needs to be recharged. Moreover, energy consumption is not only a question of battery

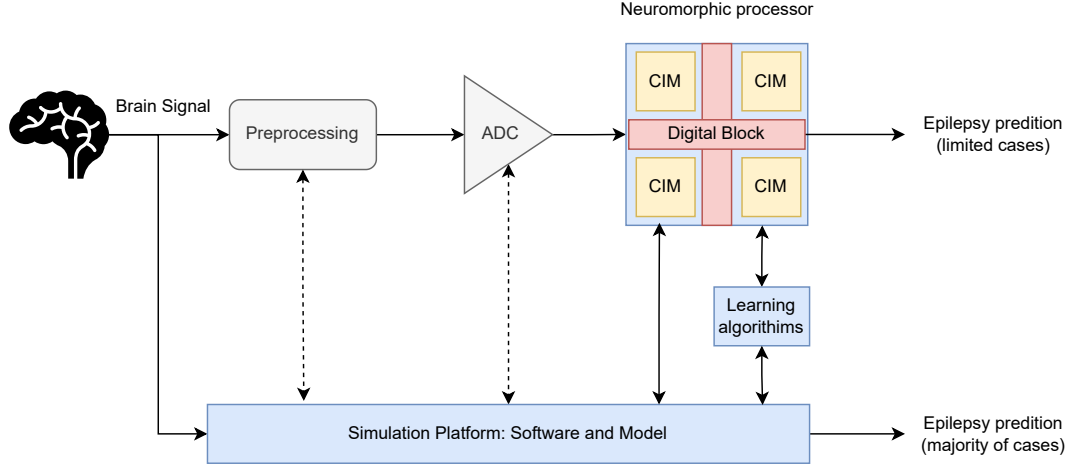


Fig. 1: High-level overview of the SELF Lab objectives

lifetime, but also a question of capacity: power-hungry models will probably be physically impossible to run in such low-powered edge devices. Such a multidisciplinary context and ubiquitous scenario raise important challenges in the way we measure and communicate energy efficiency. Developers of AI models, hardware designers, healthcare professionals, and patients might have a different way of interpreting energy efficiency, and it is important that all stakeholders have a clear understanding of it so that optimal design decisions are made.

The metrics analysis presented in this paper is done through this lab lens. While some of our conclusions can be applied to the whole field of neuromorphic computing, we will focus especially on the value that these metrics bring to healthcare edge devices and how to facilitate the development of SNN for this specific application.

IV. RELATED WORK

Despite the claims made by various papers about energy efficiency gains for SNNs, the standardization of benchmarking for SNNs' efficiency and accuracy still remains an open discussion in the community. Many papers designed custom benchmarks to test their SNN design according to the task tackled by the model. For example, SNN models that target computer vision will test their models on popular datasets like N-MNIST [19], a neuromorphic adaptation of the traditional MNIST dataset for computer vision. Then, they report energy consumption through different means, either by directly measuring on-chip usage – mostly papers that focus on hardware design – or by defining their own approximations based on performed operations or the complexity of the model.

The work by Ostrau et al. [20] introduces a set of benchmarks for open-loop tasks – the system does not adapt its behaviour based on the output – and compares the performance to the human brain based on individual costs of certain actions of the brain, and mapping those actions to neuromorphic chip operations. However, these benchmarks still focus heavily on

hardware rather than SNN model design, and lack formal definition or standardization. Other tasks show a lack of benchmarking options, like closed-loop tasks – the system acts based on predictions, e.g. detect an epilepsy attack and apply neurostimulation to avoid it – as pointed by Milde et al. [21]

NeuroBench [1] is the first attempt at providing a standardized benchmarking framework, built on top of Pytorch and SNNtorch, in collaboration with a large part of the neuromorphic community. They provide multiple standardized datasets for different tasks, like computer vision or speech processing.

The framework facilitates the computation of a series of metrics for measuring both accuracy and estimating energy and resource usage of the model. For representing energy usage of the model, the metrics provided focus on the number of synaptic operations (MAC and AC operations) and the number of membrane updates as main proxies to energy consumption.

However, later works agree that only measuring the number of operations is not enough to accurately estimate energy usage, and other metrics should be considered. Hueber et al. [3] extend this benchmarking for brain-computer interfaces, using NeuroBench but considering memory footprint and number of memory accesses as additional metrics for estimating power consumption. However, they derive the number of memory accesses from the number of operations computed by NeuroBench, estimating three loads and one store for each MAC operation, and two loads and one store for each AC operation.

On the other hand, Lemaire et al. [2] also consider the MAC, AC, and memory accesses metrics, but propose an analytical estimation rather than runtime measurements, derived from the architecture of the model. They provide a series of equations to compute total operations and accesses for each layer of the SNN model and estimate energy consumption by assigning an energy cost to each type of operation and access, based on

previous literature.

V. METHODOLOGY

In this Section, we introduced the methodology followed for our review. Figure 2 shows an overview of this methodology, which consists of three major steps. First, we collect relevant papers on SNN benchmarking and energy metrics by applying snowballing from the NeuroBench paper. Secondly, we classify the metrics according to 4 properties: Accessibility, Fidelity, Actionability, and Trend-Based. Finally, we incorporate feedback from practitioners from the SELF Lab, who provide additional metrics and papers.

A. Data collection

While research on neuromorphic computing and SNNs has grown considerably in the last decade, works that focus on benchmarking energy efficiency are still relatively niche [22, 7]. Therefore, the benchmarking literature is still not extensive, and papers that focus on energy measurement and estimation in this field are rare. For this reason, we adopt an informal literature review to explore existing energy metrics in this context.

As pointed out in Section IV, the standard observed in research is to create custom benchmarks to measure accuracy and energy efficiency. NeuroBench [1] was the first attempt to standardize the energy efficiency benchmarks for SNNs. However, later works criticized and proposed new metrics and methodologies.

We start by analyzing the metrics proposed in NeuroBench, and collect related work by performing forward and backward snowballing in the original NeuroBench paper [1], manually checking citations in NeuroBench and papers that cite NeuroBench. We used Google Scholar to gather forward and backward references available as of January 2025. Due to the recent nature of this topic, we have also considered preprints that are not yet peer-reviewed. From this first selection, we only kept papers that focus on benchmarking of SNNs and discussions on energy metrics. Afterwards, we discussed with practitioners from the SELF Lab, who pointed to additional papers and metrics. In total, we analyzed 111 papers and found 9 papers that presented energy metrics for neuromorphic AI solutions.

B. Classification

In this section, we introduce our classification of the different energy performance metrics used in neuromorphic literature. We will analyze this from the Software Engineering perspective. This means that we will study the metrics from the perspective of an SNN model developer. We consider that this professional has knowledge of the theoretical foundations of SNNs and how to develop a working model that performs a specific task. However, the developer mostly works on traditional hardware (e.g., a work laptop) and might not have extensive knowledge about neuromorphic hardware, as well as limited access to the hardware where the SNN will be deployed.

In total, we collect 13 energy metrics applicable to the implantable neuromorphic field. We analyze these metrics based

on four properties that we consider relevant in our context: Accessibility, Fidelity, Actionability, and Trend-Based. We pinpoint these different properties below.

1) *Accessibility*: In this work, we define a metric as having *Accessibility* if it can be easily obtained by an SNN developer while working on their model without the need to access hardware. In the context of neuromorphic computing, this would be a metric that is computed based on the model execution and does not require access to neuromorphic hardware to be obtained.

As highlighted previously, an SNN developer will probably have limited access to the target hardware and might require assistance from the hardware designers in case of more experimental projects. Accessible metrics can help detect energy inefficiencies in the model earlier, which can be ironed out before deploying to hardware, saving time and resources.

2) *Fidelity*: We say an energy metric has high *Fidelity* if it accurately reflects the real energy usage of the model in production. A metric with a high fidelity would be on-chip power measurements, but they are also the least accessible. Therefore, we also consider other approximations of real energy usage that have been validated by research and industry. For example, during the development of a new hardware architecture, simulation software is typically used to build the chip and test energy consumption, among other functionalities. While these results might not be exactly those observed with real-world measurements once the hardware is built, it is considered a pretty accurate approximation.

On the other hand, metrics like the number of operations do not have high fidelity. While the total energy consumption of an SNN is directly related to the number of operations performed, the real energy usage will heavily depend on the hardware running the model.

3) *Actionability*: We say a metric provides *Actionability* if it helps an SNN developer to decide if they should work on improving the energy efficiency of the model under development.

According to the findings by Ram et al. [23], actionable metrics must be practical, contextual, and exhibit high data quality (defined as accurate, consistent, and credible).

In our context, an actionable metric is practical if it is easy to interpret for the developer, and ideally, it reflects approachable concepts, rather than theoretical and non-intuitive measurements. For example, the total number of MAC operations might reflect total energy, but it can be difficult to gauge if optimizations are necessary based only on that number.

While many of the metrics found in the literature reflect energy efficiency with different degrees of accuracy, many of them are not actionable.

4) *Trend-based*: Trend-based metrics are those that can indicate improvement or degradation of energy performance based on previous measurements. These are metrics that, while not actionable per se, can give a sense of how changes to the model might affect energy consumption during deployment.

An example of this metric would be the number of parameters or the number of synaptic operations metrics mentioned previously. A single measurement of the number of operations

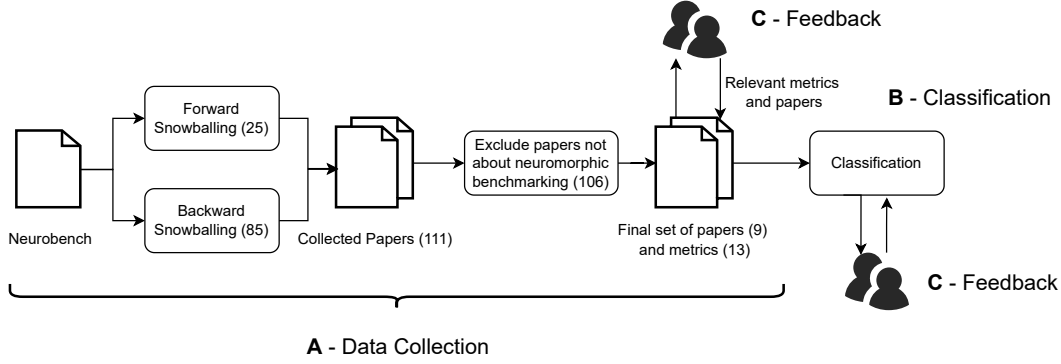


Fig. 2: Methodology followed for finding and reviewing the metrics

required for an inference does not relay information about final energy costs. However, if a developer finds out that this number increases after making changes to the model, it would be a fair conclusion

C. Feedback from practitioners

To mitigate author biases in the classification of energy metrics, we ask three members of the SELF Lab to review the classification and provide their own perspectives on the presented metrics.

The team at the SELF Lab is multidisciplinary, with practitioners who handle hardware design and SNN model design. One member works on mixed analog-digital neuromorphic hardware design and architecture. Another member specializes in embedded and implantable Machine Learning devices. Finally, the third member works on SNN model and algorithm design.

We asked them to provide their opinions on the properties defined, and asked them to classify the metrics we collected using the properties blindly, without access to our classification. With this, we confirmed our classification and updated some metrics' properties based on the posterior discussion. The practitioners also provided some additional papers that proposed metrics to look into.

VI. RESULTS

Following our methodology, we collect 13 metrics that are used to assess energy efficiency in neuromorphic software systems. Table I outlines each metric, providing a short description, our classification based on Accessibility, High Fidelity, Actionability, and Trend-Based properties, and references to neuromorphic literature that use the respective metric.

The "Reference" column indicates the paper or papers that discuss the metric as a metric for energy consumption in SNNs. When accompanied with a P, it indicates that the metric was proposed by the practitioners.

Below, we examine each property in detail, analyzing how the metrics contribute to representing and informing developers about energy efficiency.

A. Accessibility

We identify a large group of metrics (8 out of 13) that have high Accessibility. These metrics are, among others, the number of Parameters, Effective Synaptic operations, or Memory Accesses. The main sources that mention or define these metrics are those works that focus on energy reporting and benchmarking [1, 2, 3].

The reason for the high Accessibility of these metrics is that they can be obtained directly from the architecture of the model, or by running the model in non-neuromorphic hardware, while keeping track of the operations that take place. There is no need to run the model on neuromorphic hardware at the development stages since the values reported should be the same regardless of the hardware.

In contrast, we observe another set of metrics that are not Accessible – 5 out of 13. These metrics were obtained from literature that focuses on hardware design and testing [26, 28, 27] and are common metrics to report results and compare performance to other architectures.

The low Accessibility of these metrics derives from the fact that they are obtained by directly measuring the hardware with power meters or other equipment, or through simulation tools. Direct measurements by a model developer are usually not feasible because of the limited access to neuromorphic hardware. For example, the SELF Lab only has one chip of this kind, and other popular neuromorphic architectures like Intel Loihi are only accessible through remote connection, making hardware measurements difficult to obtain. On the other hand, according to the practitioners consulted, simulation tools in this field are usually closed-source, like Cadence Virtuoso – a suite of tools for hardware design and simulation –, have a steep learning curve, and are hard to integrate with other tools, making their use hard for people not specialized in hardware design.

A common metric found across neuromorphic literature [26, 27] is Energy per Learning and Energy per Inference, which represents the energy necessary, in Joules, to process a single training sample or inference. These are useful metrics for comparing the performance and efficiency of different hardware architecture proposals. However, these metrics, when measured

TABLE I: Metrics table with descriptions and attributes.

Metric	Description	Accessibility	High Fidelity	Actionability	Trend-Based	Reference
Parameters	Trainable and/or non-trainable parameters. More can mean more energy.	Yes	No	No	No	[1]
Effective Synaptic Operations	MACs and ACs in the synaptic connections of the network.	Yes	No	No	Yes	[1, 2, 20]
Membrane Updates	MACs and ACs operations performed to update the potential of the neurons.	Yes	No	No	Yes	[1, 2, 20]
Activation Sparsity	Density of 0s across the whole network for an inference or group of inferences.	Yes	No	Yes	Yes	[1]
Memory Footprint	Size of the model in memory.	Yes	No	No	No	[3, 1]
Connection Sparsity	Connections between layers.	Yes	No	No	No	[1]
Memory Accesses	Read and write operations in memory.	Yes	No	No	Yes	[3, 2]
Training Time	Total time taken to train the model.	Yes	No	No	Yes	[24, 25]
Energy per Inference	Energy consumption of performing a single inference task. Used in SoC ML.	No	Yes	No	No	[26, 27]
Energy per Learning	Energy consumption of processing a training sample.	No	Yes	No	No	[27]
Energy Area FoM	Considers power per channel, area, and sampling frequency. Relatively novel, not used.	No	Yes	No	No	[28]
Peak per Energy Consumption	Energy consumed per system operation (SOP) at peak performance, typically expressed in picojoules per system operation (pJ/SOP).	No	Yes	No	No	P [27]
Power Density	Power delivery per area of the chip (mW/cm^2). Safe limits are imposed based on the functionality of the device.	No	Yes	Yes	No	P [29]

on the hardware, do not give insights into the model itself, since hardware energy usage also depends on other components or power leakage from the architecture.

Other generic metrics for neuromorphic implantable devices that do not necessarily focus on AI performance are Peak per Energy Consumption, which represents energy consumption of individual system operations, or Energy-Area Figure of Merit (E-A FoM). This last metric is relatively novel, proposed by Zhu et al. [28] in the field of implantable chips. This metric combines energy and chip sizes to introduce form factors into energy efficiency measurements. However, this metric has still seen limited adoption.

B. Fidelity

When analyzing the metrics from the Fidelity perspective, we find that 8 out of 13 metrics present Low Fidelity. We also see that this has an opposite relationship to Accessibility – all of these metrics with Low Fidelity are previously identified as Accessible.

Metrics that have low fidelity come from software benchmarking of the SNN, and, for most of them, a proportional relationship to energy usage can be assumed. However, deriving an accurate estimation of energy consumption is difficult, since it will depend on other parameters, like the target hardware architecture and the energy costs of operations in this hardware.

An example of a metric that presents this proportional relationship with energy is the number of Effective Synaptic Operations that appears proposed in various papers [1, 2]. This metric counts the number of MAC and AC operations taking place during inference in the connections between neurons, and it is similar to the floating point operations metric (FLOPs) used

in Green AI literature as a proxy to energy consumption [30]. However, this metric alone does not provide enough information to obtain an accurate estimate of energy consumption, since detailed information about the target hardware is required [31].

On the other hand, the high-fidelity metrics we identified – 5 out of 13 – are also not Accessible. These metrics are obtained through direct experimentation in the hardware with power meters or through detailed hardware simulations. In the reviewed literature, examples of the latter are much more abundant. For example, in the work by Yaldagard et al. [17], a neuromorphic chip design is loaded into the Cadence Virtuoso Suite, given a set of tasks for benchmarking, and electric currents across the chip are simulated.

The energy measurements obtained from these methods accurately represent the energy consumption. Readings from a power meter directly represent energy consumption, while measurements from simulation tools are precise enough to be considered valid results for scientific literature, being used across the reviewed papers [17, 28].

C. Actionability

Looking into the metrics that show the Actionability property, we observe that only two of the collected fulfill our definition of Actionable. We identified these metrics as Actionable after our discussion with practitioners of the SELF Lab.

The first Actionable metric is Activation Sparsity, which represents the density of zeros across the network when performing inferences. A higher sparsity means fewer calculations in the network, and efficiency is higher. The reason to consider this metric as Actionable is that, in the field of SNN research, a common rule of thumb is that a model with sparsity lower

than 60% is considered an inefficient model that does not fully exploit the capabilities of neuromorphic computing. The discussion with the practitioners from the SELF Lab confirmed this rule of thumb. Therefore, if a developer observes an Activation Sparsity lower than 60%, they know they must take action to improve the design.

The second identified Actionable metric is Power Density. The reason for its Actionability is that health organizations, like the Food and Drug Administration (FDA) in the United States, set maximum safe limits on power delivery for implantable devices based on their functionality. For example, for RF-emitting medical devices, the power density is limited to 10 mW/cm². If the power density metric is above the limits set for the target application, the developer knows that they must change their design.

D. Trend-Based

While the majority of the identified metrics are not Actionable, 5 out of the 11 non-Actionable metrics are Trend-Based. Trend-Based metrics can give a sense of how changes to the model might affect energy consumption during deployment, giving some Actionability if they are actively monitored during development. The identified Trend-Based metrics are Effective Synaptic Operations, Membrane Updates, Activation Sparsity, Memory Accesses, and Training Time, which are also Accessible.

The reason these metrics are considered Trend-Based is that a proportional relationship to energy usage can be assumed, even if they cannot fully represent energy consumption for the whole model.

In a context of implantable devices, where we have a single, well-defined target architecture, we can safely assume that a higher number of Effective Synaptic Operations will translate into higher energy usage. Keeping a record of these metrics across different model versions can help the developer assess how they compare in terms of energy efficiency.

VII. DISCUSSION

In this section, we identify two key challenges arising from our studies: the lack of metrics that are both accessible and high fidelity, and the lack of actionable metrics. To address these issues, we provide recommendations for potential research directions. Additionally, we discuss the threats to validity and limitations of our study.

A. Bridging Accessibility and Fidelity

The first challenge we can identify is the lack of metrics that are both Accessible and have High Fidelity. This comes from the fact that literature on SNN benchmarking mostly focuses on the software side, measuring metrics that stem from SNN model design, and they fail to consider the impact of the target hardware. This is a highly relevant challenge, since the lack of Accessible metrics that can accurately reflect the energy consumption of an SNN model means that obtaining information about the energy efficiency of a model is not possible until a very late phase of the project.

A recommendation to solve this problem is to build energy estimation methods for SNN models, considering specifications

about the chip provided by the hardware designers. This estimation should not only account for the energy usage of the model, but also the energy usage of other features of the device, like pre- and postprocessing or wireless data transmission.

According to the practitioners we consulted, some of the specifications that could be provided from simulations are energy costs of MAC and AC operations, energy costs of memory accesses, or energy costs of processing in between layers, for mixed analog-digital neuromorphic designs. Other specifications for features unrelated to the model could be the CMOS architecture used for chip fabrication, which determines standby power leakage. For certain applications, other costs have to be considered. For example, for a brain implantable device, an Analog-Digital Converter is necessary to convert the brain waves readings to digital data for the model.

For this estimation to be useful, it should be divided into costs that come purely from the SNN model and costs that come from the other elements of the chip. Depending on the target application for the hardware design, the difference in energy and power consumption of the model itself can be several orders of magnitude smaller than the power used by other systems [32]. Therefore, in some cases, it might not be useful to keep improving model performance for a negligible impact on total energy consumption.

B. Lack of Actionable Metrics

A second challenge identified is the lack of actionable metrics across the literature. This lack of actionable metrics impedes the developer from making informed decisions about the energy performance of the model. Actionable metrics during model design can provide early feedback for SNN models in projects that practice hardware and software co-design. This early feedback for software is highly relevant, since iterating over different software designs is much faster than iterating and improving the hardware architecture. For example, in mixed analog-digital neuromorphic hardware, the analog part of the hardware is not programmable, and its architecture partly depends on the model. Therefore, it is important to know as much information about the energy efficiency of the model as possible in the early phases of development.

To tackle this challenge, we present four recommendations about possible research directions for actionable. These focus on assessing metrics with inherent Trend-Based properties from Sustainable Software Engineering, metrics that detect large changes in power requirements, metrics for early evaluation of battery life, and metrics that measure energy-performance tradeoff.

Table II presents some of the metrics that could come out of these research directions, evaluated according to our classification methodology. We update the Accessibility and High Fidelity field, and include a * if the metric is both Accessible and with High Fidelity, assuming that a reliable energy estimation is built as mentioned previously.

1) *Extending Trend-Based metrics*: Our review uncovered that while many metrics provide information about the energy consumption of an SNN model, not many of them do so in an

TABLE II: Overview of potential research directions to address challenges in existing literature

Metric	Description	Accessibility	High Fidelity	Actionability	Trend-Based	Reference
Energy Delay Product	Product of energy consumption and execution time	Yes*	Yes*	No	No	NA
Speedup	Time _{new} /Time _{old} , change of execution time after changes	Yes	Yes	Yes	Inherent	[33]
Greenup	Energy _{new} /Energy _{old} , change of total energy after changes	Yes*	Yes*	Yes	Inherent	[33]
Powerup	Speedup/Greenup. If greater than 1, new code uses more power on average	Yes*	Yes*	Yes	Inherent	[33]
Estimated battery life	Expected battery life given energy per inference and battery capacity	Yes*	Yes*	Yes	No	NA
Inferences per battery cycle	Inferences that can be done with a full charge. Opposite of the previous one	Yes*	Yes*	Yes	No	NA
Accuracy-Efficiency Tradeoff	Energy costs against accuracy gains between versions of the model	Yes*	Yes*	Yes	Yes	[34]

* assuming energy estimation

Actionable way. However, Trend-Based metrics can provide certain Actionability when actively keeping track of them.

To tackle this problem, we propose incorporating some metrics from the Sustainable Software Engineering field that are inherently Trend-Based and can prove useful in the context of neuromorphic computing in general and implantable neuromorphic devices. Some of these metrics are derived from energy usage or execution time, which means they directly depend on the energy estimation previously mentioned.

Some useful metrics during model development are Speedup, Greenup, and Powerup [33], which can indicate improvements in performance and energy efficiency across the development of the model. Therefore, they are Trend-Based by definition, in the sense that they are already derived from historical data.

2) *Large Changes in Power Requirements*: In the context of implantable neuromorphic devices, power delivery is highly limited by practical and medical reasons. This is highlighted by metrics like power density, for which health organizations establish safe maximum limits. Therefore, metrics that easily indicate large changes in power requirements for the model are crucial for the developer.

The previously mentioned Powerup or the Energy Delay Product metric can be applied in this context. These metrics combine execution time and energy to give a sense of total power usage. This is extremely relevant for implantable brain devices, due to the power per area limits set for these applications. High power delivery and/or long execution times of the model can lead to elevated temperatures, which can be uncomfortable or damaging to the patient.

3) *Battery related metrics*: Another aspect of the model that cannot be quickly evaluated from the reviewed metrics is the estimated battery life of the device. This is a highly relevant problem in the field of implantable brain devices, since batteries are implanted together with the chip through brain surgery, and cannot be recharged or replaced without further surgery. Therefore, according to practitioners, a common quality attribute used in this field is that batteries should last for at least 10 years, to minimize the number of invasive brain surgeries performed as much as possible.

To cover this aspect of SNN development for implantable devices, we recommend research directions for defining and estimating metrics like Expected Battery Life or Expected

Number of Inferences per Battery Cycle. These metrics can be derived from an energy estimation based on specifications, battery capacity, and consumption of other components. With them, developers can access an easy overview of the impact of the model on battery life, and making decisions that target energy efficiency is easier.

4) *Assessing Energy-Accuracy Tradeoff*: Another relevant aspect of model development that lacks Actionable metrics for evaluation is assessing the tradeoff between energy and accuracy of the model. The traditional development pipeline of an AI model focuses on iterating designs to maximize accuracy. However, sometimes small improvements in accuracy can come at a high energy cost. In the implantable field, where power and energy are highly limited, sacrificing accuracy slightly could be worth it compared to the energy efficiency gains.

Possible research directions to solve this problem could be defining a metric or set of metrics that accurately represent the Energy-Accuracy tradeoff between model versions. This metric would be useful when iterating over different versions of the model. Estimations of energy combined with the reported accuracy of the model can be used to report an increase in energy consumption against gained accuracy and help determine if potential accuracy increases are worth the additional energy cost.

A previous work by Cueto-Mendoza and Kelleher [34] defines a similar Efficiency Ratio metric for AI models, defined as *Accuracy/Energy*. This definition is limited, since a model with good accuracy and a model with poor accuracy could show similar values if the second one uses much less energy.

Therefore, it is not a very actionable metric. As an example, suppose we have a version V1 of a model that reaches 0.7 accuracy with a certain energy cost per inference, and a version V2 that reaches 0.8 using double the amount of energy. Efficiency Ratio, as defined, would not give much insight into this breakdown. A more sophisticated metric would help determine if it is worth doubling energy consumption in exchange for this additional 0.1 accuracy. In the implantable context, it could even be combined with the previously mentioned battery-related metrics to determine if the hit to battery life would be significant.

C. Threats to Validity

The nature of this study presents some limitations. Firstly, we focus our scope on SNNs, and specifically on implantable healthcare devices. Therefore, our findings might not generalize to other domains of AI. In general, we believe most of our reflections could be applied to the field of edge AI, given the similarities between both fields. Edge AI also deals with low-power devices and sometimes with devices running on a battery. Additionally, some reflections could be applied to the field of Green AI in general, since these metrics focus on reporting energy consumption. However, the definition and validity of these metrics for other fields or applications would have to be evaluated independently.

The lens of the project through which we study the problem also limits the feedback from the practitioners. These practitioners come from a concrete neuromorphic project from the SELF Lab. Their views are shaped by this project, and some conclusions might not apply to other projects.

The limited research done in SNN benchmarking is also a limiting factor for the collection methodology applied for this review. The research and first attempts at benchmark standardization started as recently as 2023. Therefore, there is little literature to apply a systematic method, so we performed an ad hoc literature review. To mitigate problems derived from this method, we applied snowballing over the Neurobench paper [1] to cover a larger number of relevant papers. Additionally, there could be new research that tackles these problems in the near future.

VIII. CONCLUSION

Neuromorphic computing and SNNs are a promising new paradigm for improving the energy efficiency of Artificial Intelligence models. However, as highlighted in this study, the research has focused mainly on evaluating hardware and models in conjunction, while research into only model benchmarking is relatively new.

By looking into the works that focus on SNN benchmarking, we uncovered and categorized the most common metrics for assessing the energy consumption of an SNN model, based on Accessibility, Fidelity, and Actionability. With this classification, we identified which metrics approximate energy consumption better and are easier to measure during model development.

We find out that there are no energy metrics that are both Accessible and present High Fidelity. Benchmarking frameworks for SNNs focus mainly on model design and metrics, without considering the effects of target hardware. This makes early evaluation of the model difficult, and we recommend research directions towards building reliable estimations that bridge Accessibility and Fidelity.

The study also highlighted the lack of actionable metrics for developers. Despite the variety of metrics, none of them focused on giving the developer early and approachable feedback. Therefore, from the lenses of implantable neuromorphic devices, we proposed research directions to extend the catalog with actionable metrics, focusing on accurately and transparently representing energy and power trends during development, estimated battery life, and Energy-Accuracy tradeoff.

By identifying gaps in existing energy metrics for SNNs in neuromorphic implants and proposing research directions to bridge them, this study lays the groundwork for more effective benchmarking methods. Our recommendations open future research into energy metrics that provide quicker feedback and better Actionability, ultimately supporting developers in designing more efficient and practical SNN models.

ACKNOWLEDGMENTS

Our research is partially funded and aimed to support the work by the Smart Edge Lab for Healthcare (SELF Lab) from TU Delft.³

REFERENCES

- [1] J. Yik, K. V. den Berghe, D. den Blanken, Y. Bouhadjar *et al.*, “Neurobench: A framework for benchmarking neuromorphic computing algorithms and systems,” 2025. [Online]. Available: <https://arxiv.org/abs/2304.04640>
- [2] E. Lemaire, L. Cordone, A. Castagnetti, P.-E. Novac *et al.*, “An analytical estimation of spiking neural networks energy efficiency,” in *Neural Information Processing*, M. Tanveer, S. Agarwal, S. Ozawa, A. Ekbal, and A. Jatowt, Eds. Cham: Springer International Publishing, 2023, pp. 574–587.
- [3] P. Hueber, G. Tang, M. Sifalakis, H.-P. Liaw *et al.*, “Benchmarking of hardware-efficient real-time neural decoding in brain–computer interfaces,” *Neuromorphic Computing and Engineering*, vol. 4, no. 2, p. 024008, 2024.
- [4] N. Tazi, F. Mom, H. Zhao, P. Nguyen, M. Mekouri, L. Werra, and T. Wolf, *The Ultra-Scale Playbook: Training LLMs on GPU Clusters*. Hugging Face, February 2025. [Online]. Available: <https://huggingface.co/spaces/nanotron/ultrascale-playbook>
- [5] W. Maass, “Networks of spiking neurons: The third generation of neural network models,” *Neural Networks*, vol. 10, no. 9, pp. 1659–1671, 1997. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0893608097000117>
- [6] N. Brunel and M. C. W. van Rossum, “Lapicque’s 1907 paper: from frogs to integrate-and-fire,” *Biological Cybernetics*, vol. 97, no. 5, pp. 337–339, 2007.
- [7] C. D. Schuman, S. R. Kulkarni, M. Parsa, J. P. Mitchell *et al.*, “Opportunities for neuromorphic computing algorithms and applications,” *Nature Computational Science*, vol. 2, no. 1, pp. 10–19, 2022. [Online]. Available: <https://doi.org/10.1038/s43588-021-00184-y>
- [8] A. Paszke, S. Gross, F. Massa, A. Lerer *et al.*, “Pytorch: An imperative style, high-performance deep learning library,” *CoRR*, vol. abs/1912.01703, 2019. [Online]. Available: <http://arxiv.org/abs/1912.01703>
- [9] J. K. Eshraghian, M. Ward, E. O. Neftci, X. Wang *et al.*, “Training spiking neural networks using lessons from deep learning,” *Proceedings of the IEEE*, vol. 111, no. 9, pp. 1016–1054, 2023.

³<https://www.tudelft.nl/ai/self-lab>

- [10] W. Fang, Y. Chen, J. Ding, Z. Yu *et al.*, “Spikingjelly: An open-source machine learning infrastructure platform for spike-based intelligence,” *Science Advances*, vol. 9, no. 40, p. eadi1480, 2023. [Online]. Available: <https://www.science.org/doi/abs/10.1126/sciadv.adi1480>
- [11] B. Yin, F. Corradi, and S. M. Bohté, “Effective and efficient computation with multiple-timescale spiking recurrent neural networks,” in *International Conference on Neuromorphic Systems 2020*, ser. ICONS 2020. New York, NY, USA: Association for Computing Machinery, 2020. [Online]. Available: <https://doi.org/10.1145/3407197.3407225>
- [12] S. Höppner, Y. Yan, A. Dixius, S. Scholze *et al.*, “The spinnaker 2 processing element architecture for hybrid digital neuromorphic computing,” *CoRR*, vol. abs/2103.08392, 2021. [Online]. Available: <https://arxiv.org/abs/2103.08392>
- [13] M. Davies, N. Srinivasa, T.-H. Lin, G. Chinya *et al.*, “Loihi: A neuromorphic manycore processor with on-chip learning,” *IEEE Micro*, vol. 38, no. 1, pp. 82–99, 2018.
- [14] J. Göltz, L. Kriener, A. Baumbach, S. Billaudelle, O. Breitwieser *et al.*, “Fast and energy-efficient neuromorphic deep learning with first-spike times,” *Nature machine intelligence*, vol. 3, no. 9, pp. 823–835, 2021.
- [15] K. N. Khan, M. Hirki, T. Niemi, J. K. Nurminen *et al.*, “Rapl in action: Experiences in using rapl for power measurements,” *ACM Trans. Model. Perform. Eval. Comput. Syst.*, vol. 3, no. 2, mar 2018. [Online]. Available: <https://doi.org/10.1145/3177754>
- [16] V. J. Reddi, C. Cheng, D. Kanter, P. Mattson *et al.*, “Mlperf inference benchmark,” in *2020 ACM/IEEE 47th Annual International Symposium on Computer Architecture (ISCA)*, 2020, pp. 446–459.
- [17] M. A. Yaldagard, S. Diware, R. V. Joshi, S. Hamdioui, and R. Bishnoi, “Read-disturb detection methodology for rram-based computation-in-memory architecture,” in *2023 IEEE 5th International Conference on Artificial Intelligence Circuits and Systems (AICAS)*, 2023, pp. 1–5.
- [18] C. Zhang, W. Pan, and C. D. Santina, “Nisnn-a: Non-iterative spiking neural networks with attention with application to motor imagery eeg classification,” 2025. [Online]. Available: <https://arxiv.org/abs/2312.05643>
- [19] G. Orchard, A. Jayawant, G. K. Cohen, and N. Thakor, “Converting static image datasets to spiking neuromorphic datasets using saccades,” *Frontiers in Neuroscience*, vol. 9, 2015. [Online]. Available: <https://www.frontiersin.org/journals/neuroscience/articles/10.3389/fnins.2015.00437>
- [20] C. Ostrau, C. Klarhorst, M. Thies, and U. Rückert, “Benchmarking neuromorphic hardware and its energy expenditure,” *Frontiers in Neuroscience*, vol. 16, 2022. [Online]. Available: <https://www.frontiersin.org/journals/neuroscience/articles/10.3389/fnins.2022.873935>
- [21] M. B. Milde, S. Afshar, Y. Xu, A. Marcireau *et al.*, “Neuromorphic engineering needs closed-loop benchmarks,” *Frontiers in Neuroscience*, vol. 16, 2022. [Online]. Available: <https://www.frontiersin.org/journals/neuroscience/articles/10.3389/fnins.2022.813555>
- [22] N. J. Pritchard, A. Wicenc, M. Bennamoun, and R. Dodson, “A bibliometric review of neuromorphic computing and spiking neural networks,” 2023. [Online]. Available: <https://arxiv.org/abs/2304.06897>
- [23] P. Ram, P. Rodríguez, M. Oivo, S. Martínez-Fernández *et al.*, “Actionable software metrics: An industrial perspective,” in *Proceedings of the 24th International Conference on Evaluation and Assessment in Software Engineering*, ser. EASE ’20. New York, NY, USA: Association for Computing Machinery, 2020, p. 240–249. [Online]. Available: <https://doi.org/10.1145/3383219.3383244>
- [24] S. Karilanova, M. Fabre, E. Neftci, and A. Özçelikkale, “Zero-shot temporal resolution domain adaptation for spiking neural networks,” *arXiv preprint arXiv:2411.04760*, 2024.
- [25] S. R. Kulkarni, M. Parsa, J. P. Mitchell, and C. D. Schuman, “Benchmarking the performance of neuromorphic and spiking neural network simulators,” *Neurocomputing*, vol. 447, pp. 145–160, 2021. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0925231221003969>
- [26] W. Ke, M. Khoei, and D. Muir, “Neurobench: Dcase 2020 acoustic scene classification benchmark on xyloaudio 2,” *arXiv preprint arXiv:2410.23776*, 2024.
- [27] F. Chen, Q. Tian, L. Xie, Y. Zhou, Z. Wu *et al.*, “Epoc: A 28-nm 5.3 pj/sop event-driven parallel neuromorphic hardware with neuromodulation-based online learning,” *IEEE Transactions on Biomedical Circuits and Systems*, 2024.
- [28] B. Zhu, U. Shin, and M. Shoaran, “Closed-loop neural prostheses with on-chip intelligence: A review and a low-latency machine learning model for brain state detection,” *IEEE Transactions on Biomedical Circuits and Systems*, vol. 15, no. 5, pp. 877–897, 2021.
- [29] N. Ahmadi-Dastgerdi, H. Hosseini-Nejad, and H. Alinejad-Rokny, “A hardware-efficient on-implant spike compression processor based on vq-dae for brain-implantable microsystems,” *Medical & Biological Engineering & Computing*, 2025. [Online]. Available: <https://doi.org/10.1007/s11517-025-03317-x>
- [30] R. Desislavov, F. Martínez-Plumed, and J. Hernández-Orallo, “Trends in ai inference energy consumption: Beyond the performance-vs-parameter laws of deep learning,” *Sustainable Computing: Informatics and Systems*, vol. 38, p. 100857, 2023. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S2210537923000124>
- [31] A. Asperti, D. Evangelista, and M. Marzolla, “Dissecting flops along input dimensions for greenai cost estimations,” in *Machine Learning, Optimization, and Data Science*, G. Nicosia, V. Ojha, E. La Malfa, G. La Malfa, G. Jansen, P. M. Pardalos, G. Giuffrida, and R. Umeton, Eds. Cham: Springer International Publishing, 2022, pp. 86–100.
- [32] C.-W. Tsai, R. Jiang, L. Zhang, M. Zhang *et al.*, “Seizure-

- cluster-inception cnn (scicnn): A patient-independent epilepsy tracking soc with 0-shot-retraining,” *IEEE Transactions on Biomedical Circuits and Systems*, vol. PP, pp. 1–12, 10 2023.
- [33] S. Abdulsalam, Z. Zong, Q. Gu, and M. Qiu, “Using the greenup, powerup, and speedup metrics to evaluate software energy efficiency,” in *2015 Sixth International Green and Sustainable Computing Conference (IGSC)*, 2015, pp. 1–8.
- [34] E. Cueto-Mendoza and J. Kelleher, “A framework for measuring the training efficiency of a neural architecture,” *Artificial Intelligence Review*, vol. 57, no. 12, p. 349, 2024. [Online]. Available: <https://doi.org/10.1007/s10462-024-10943-8>