Why TinyML?
Exploring the reasons why TinyML is used for real-world problems

Marco Wagner

Faculty of Engineering, Heilbronn University, Heilbronn, Germany

Abstract. Machine Learning (ML) and especially its application to cyber-physical systems is an uprising field of research. Many approaches on how to leverage the power of ML even in small devices have been published and applied in recent years, forming the field of TinyML. While TinyML has been promising several benefits such as cost-reduction, privacy and more, studies on which of these benefits are actually crucial when deciding to apply a TinyML approach to a real-world problem have been missing so far. The author of this paper argues, that without understanding the "why", the community of researchers and industrial parties may not understand the reasons of applying TinyML and hence may head into research and development directions not increasing its success. This work analysis the application of TinyML approaches and the reasons behind in recent years for three important fields of application: consumer electronics, manufacturing and automotive. It determines the distribution of TinyML applications in the named field and examines which of the bespoken benefits of TinyML were actually driving the decision to use it. Furthermore, this work investigates in cross connections between the benefits and hence points out the main combinations of benefits to the adopter.

Keywords: TinyML, embedded systems, embedded ML, machine learning, consumer electronics, automation, automotive

1 Introduction

Artificial Intelligence (AI) and especially its sub-field of Machine Learning (ML) have gained quite some attention in recent years. With fields of application stretching from generative AI creating text or even art work all the way to more technical usage such as for example computer vision [10] or classification of network traffic [7]. However, most of these approaches have been designed to work on high-performance computers or even on large-scale cloud infrastructure which are computationally heavy in memory, CPU, network bandwidth etc. [2].
While ML applications for high-performance computers are still developing further, a new field of research has been on a rise as well that aims at creating ML applications efficient enough to run on small embedded devices: TinyML. TinyML tries to achieve several benefits ranging from both the technical problem domain with topics like resource efficiency or real-time, to economic drivers such as costs or security, also to sustainability factors and reduced energy consumption. While such benefits

David C. Wyld et al. (Eds): SIPRO, EMSA, CCSEA- 2024
pp. 13-22, 2024.CS & IT - CSCP 2024 DOI: 10.5121/csit.2024.140902

are often referred to be important (see e.g. [1], [6] and [9]), no research has been conducted so far on the question which of these benefits are actually important to the researchers and practitioners applying TinyML principles in order to create new algorithms, machine learning models or ML-based embedded systems. This work tries to close this gap by providing a comprehensive and detailed analysis of several hundred recent publications in the application domains consumer electronics, manufacturing and automotive. Furthermore, it investigates on cross connections of the benefits and tries to draw conclusions in order to steer the TinyML community into the right direction.
The remainder of the paper is structured as follows: Section 2 discusses the state of the art before Section 3 introduces and explains the main research questions to be tackled as well as the underlying research methodology. In Section 4 the results of the investigations are presented and discussed, before Section 5 summarizes the paper and gives an outlook on potential future work.

2 Related Work

The adoption of machine learning approaches for small and smallest embedded devices, also known as TinyML, has gained a lot of attention in recent years. This is mainly due to the fact that both the commercial potential of embedded systems on the one hand (for instance, it is assumed that a total quantity of 30.9 billion operational IoT devices by the end of 2025 [13]) and the size and activity of the academic community in this area on the other hand are significant.

One cluster of research paper on TinyML focuses on surveying different dimensions such as the fields of applications it is used in (e.g. [1]), specific mechanisms leveraged (e.g. quantization in [16]) or dedicated problem classes to be solved (e.g. obejct detection in [4]).

Another group of publications tries to increase the adoption of TinyML by providing detailed and easy to follow explanations on how to use the algorithms developed in the community on small and large problem scenarios (e.g. [9], [11]).
Other papers name benefits or key performance indicators, but do fail in analyzing their individual importance. Examples for this category of publications are [1], [6] and [9].

Finally, one of the trends observable is the increasing number of papers discussing the potential for more sustainable embedded systems by introducing TinyML. The authors of [12] for example, investigate not only the inference phase on the device but also the development phase and the carbon dioxide footprint caused by that. In [3] applications are discussed where the usage of TinyML could be beneficial for the environment by adding smart features to pervasive devices. And finally, [15]

discusses the potential of TinyML for more sustainable food supply chain. While these papers shed some light on important areas with potentials to increase sustainability, none of is ale to give evidence that this topic is really important to those adopting TinyML in their technical systems.

As a conclusion, to the best of our knowledge, this is the first paper that not only names the expected benefits of TinyML, but also analysis their importance, both individually and in combination, and draws a comparison across different fields of application.

3 Research questions and methodology

The following chapter introduces the research questions worked on as well as the methodology followed in this paper.

3.1 Research questions

In line with the objectives of this work, and owing to identify the "why" behind using TinyML in actual applications, three research questions are tackled:

- RQ1: How popular have TinyML approaches been in the three fields of applications consumer electronics, manufacturing and automotive?
- RQ2: Which are the most crucial benefits that led to selecting TinyML for the applications?
- RQ3: Which cross connections between the benefits do exist from the perspective of an adopter of TinyML?

The first one, RQ1, tries to spotlight the overall popularity of TinyML approaches in actual usage scenarios. This is done in the three important embedded systems domains of consumer electronics, manufacturing and automotive.

The second research question, RQ2, tries to investigate the importance of different known benefits on TinyML to the community independent from each other. In order to be able to identify different levels of importance in different domains, the three fields of usage consumer electronics, manufacturing and automotive are analyzed independently.

Finally, the last research question, RQ3, investigates on the cross-connections of benefits and their importance to the engineer or researcher deciding to make use of an TinyML approach. Again, this analysis is done in all three domains of application separately, to identify not only globally visible cross-connections but also differences across consumer electronics, manufacturing and automotive.

3.2 Research methodology

This work was conducted by executing a systematic literature review followed by a quantitative analysis of the papers collected. In more detail, the following four steps were conducted:

Step 1: systematic literature review In the first step, a systematic literature review has been conducted using google scholar. The search queries used where the following:
For consumer electronics:

KEY (tinyml)
AND (consumer electronics OR smart devices)
AND PUBYEAR $\geq$ 2022
AND (LIMIT-TO (LANGUAGE, English))

For manufacturing:

KEY (tinyml)
AND (industry 4.0 OR manufacturing)
AND PUBYEAR $\geq$ 2022
AND (LIMIT-TO (LANGUAGE, English))

For automotive:

KEY (tinyml)
AND (automotive OR vehicle)
AND PUBYEAR $\geq$ 2022
AND (LIMIT-TO (LANGUAGE, English))

These queries resulted in a list of 359 hits (consumer electronics), 579 hits (manufacturing) and 780 hits (automotive). The databases for each query where collected using the tool publish or perish [8].

Step 2: filtering the results In the second step, the lists of hits of all three search queries where filtered in order to increase the quality of the dataset. The filter criteria applied where:

- C1: Studies must present new applications of TinyML approaches in the respective field. Surveys, books and pure descriptions of frameworks where not considered.
- C2: Duplicates and false positives within the dataset where excluded.

This step resulted on a reduction of examined papers as listed in Table 1.

Table 1. Hits after filtering

appl. field	hits	hits after filtering
consumer el.	359	299
manufacturing	579	428
automotive	780	457

Step 3: quantitative analysis In the final step, the three filtered datasets undergo a quantitative analysis using a standard spreadsheet tool. In this examination, the frequency of mention of six different benefits of TinyML are calculated (list of benefits derived from [1], [6] and [9]):

- B1: **sustainability (S)** using TinyML to increase the sustainability of the final solution.
- B2: **energy efficiency (E)** leveraging the potential of TinyML to safe energy.
- B3: **resource efficiency (R)** reduce the need of resources through TinyML compared to standard ML approaches.
- B4: **costs (C)** trying to reduce the overall system costs by leveraging TinyML.
- B5: **latency and real-time (L)** using TinyML to improve latency and real-time capability of the application.
- B6: **privacy and security (P)** increasing the privacy and/or security of the system using TinyML.

The analysis was conducted manually; a match was counted if the respective benefit was either named explicitly or implicitly by corespondent performance indicators (e.g. energy consumption measurements).

4 Results

In the following sub-sections the results of the study are presented by trying to answer the research questions stated in Section 3.

4.1 RQ1: Popularity of TinyML in the areas of application

As stated in Table 1 and illustrated in Figure 1, the most popular field of application is the automotive sector, followed by manufacturing and consumer electronics. Analyzing the number of hits after filtering, TinyML was leveraged in 457 publications in automotive, followed by manufacturing with 428. Well spaced out, consumer electronics was the less popular field TinyML was used in with 299 occurrences.

While it is difficult to derive direct recommendations from this figure, it might help the TinyML community already to adjust the focal point of research by either prioritizing automotive requirements or analyzing why TinyML is not as popular right now in the two other domains.

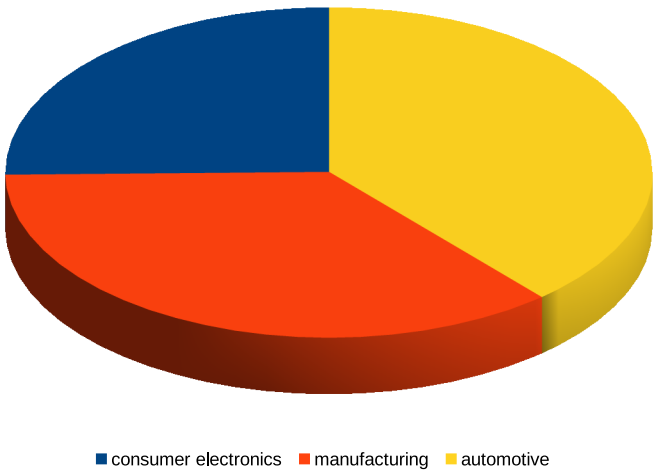

Fig. 1. Popularity within application areas

4.2 RQ2: Why TinyML: individual importance of benefits

The second research question examines the popularity of the individual benefits considered. In other words: how often have each of these benefits been named in the papers reviewed either explicitly or implicitly through quantitative indicators. While in all three domains, resource efficiency seems to play the prior role, the importance of the other benefits examined varies across the fields of usage of TinyML. The results are illustrated in Figure 3. While energy efficiency is the second most important reason to choose using TinyML for both automotive and manufacturing, it is only placed fourth in the area of manufacturing, where latency and real-time was named the second biggest reason. Somewhat surprisingly, sustainability seems not to be a real factor in any of the domains, although AI has been put to the pillory in recent years, for its huge carbon footprints (e.g. [14], [5]).

Another finding in this context is that only a small portion of the papers reviewed (only 254 out of 1184) are exclusively focused on a single benefits while the vast majority selected to use TinyML to leverage more than one benefit. Figure 2 visualizes the distribution of the number of benefits named in the examined papers and showcases that most approaches named two benefits (725 out of 1184).

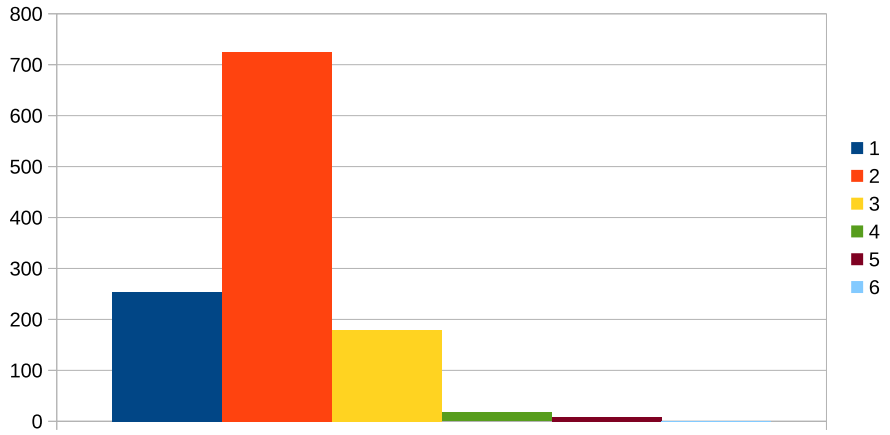

Fig. 2. Number of benefits named per paper

4.3 RQ3: Why TinyML: cross connections between the benefits

Due to the finding of the last chapter, that the majority of papers names more than one benefit to be of interest, it makes a lot of sense to not only look on the benefits as individuals, but also on the most popular combinations of benefits named by the papers surveyed. The results of this cross connection analysis are visualized in Figure 4.

One finding here is that the combination of energy efficiency & resource efficiency is very popular in all three domains of usage (most popular option in consumer electronics and automotive, second most popular in manufacturing). On the other hand, a combination like resource efficiency & costs varies in popularity between the different application areas: from being not really important in consumer electronics, to medium importance in automotive to being the top benefit combination in manufacturing. This is especially interesting, since in all three industries, cost pressure is known as a major topic.

Again, sustainability is only a very insignificant topic. This confirms the observation made in RQ2 and again puts up the question if TinyML is really able to play a noticeable role towards greener IT systems.

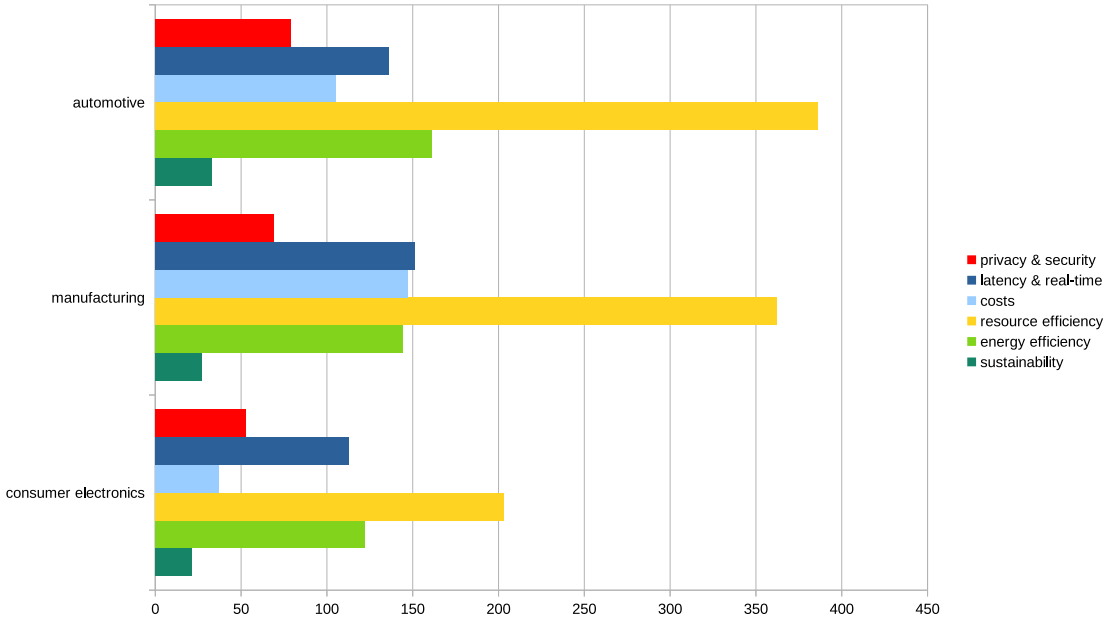

Fig. 3. Individual importance of benefits

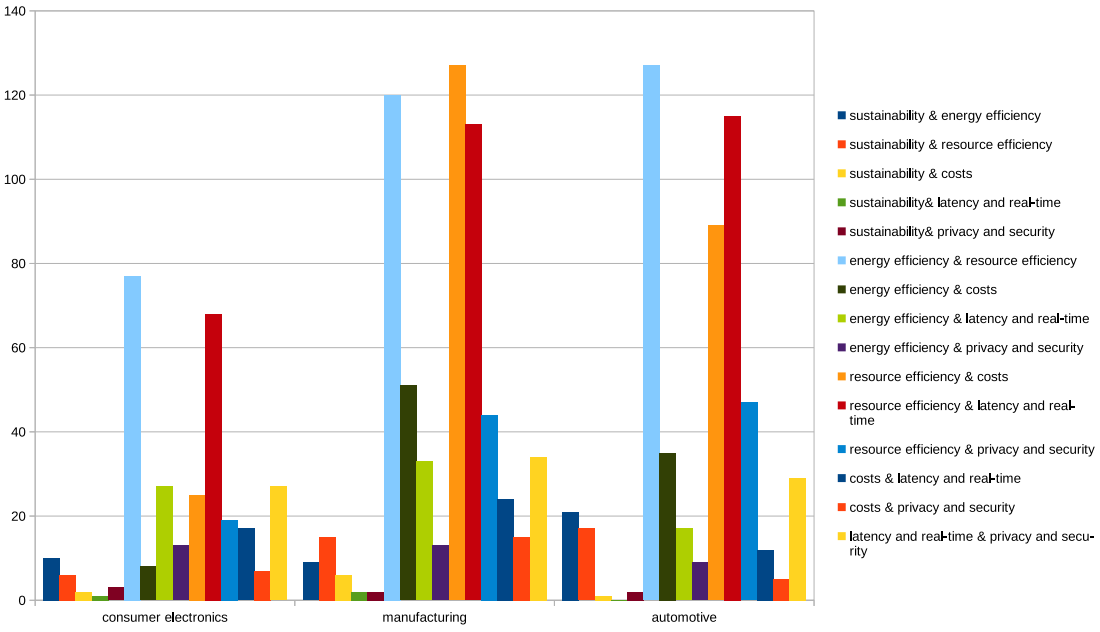

Fig. 4. Cross connections of benefits

5 Summary and outlook

In this work, the "why" behind the usage of TinyML has been investigated using three clearly defined research questions (RQ1 - RQ3). Furthermore, the underlying methodology has been described and the results of the literature study have been explained and discussed.

While in the examined publications, a clear preference towards technical benefits and especially resource efficiency could be observed, other topics, such as sustainability do not seem to be too much of interest for the community, yet.

Additionally, this study showed clearly that most of the adopters of TinyML try to leverage a combination of benefits rather than focusing on a singular once. While the total numbers vary over the domains of usage, one common finding is that again combinations of technical KPIs, such as energy efficiency & resource efficiency and resource efficiency & latency and real-time outnumber other combinations partially by factors.

While one conclusion of these results for the TinyML community could be, that it should focus even more on the technical benefits and e.g. intensify the search for algorithms with smaller footprints and execution times, these results should also lead to more efforts with regards to actually leverage the promised sustainability gains or to be honest enough to drop these claims.

Future work will enlarge this study to also look into other application domains. Furthermore, as this work only looked on very recent publications another extension of the investigation could be to compare theses results with other sections in time to potentially identify trends in the importance of benefits over time.

References

1. Abadade, Youssef, Temouden, Anas, Bamoumen, Hatim, Benamar, Nabil, Choutki, Yousra, and Senhaji Hafid, Abdelhakim. A Comprehensive Survey on TinyML | IEEE Journals & Magazine | IEEE Xplore.
2. Mattia Antonini, Miguel Pincheira, Massimo Vecchio, and Fabio Antonelli. An Adaptable and Unsupervised TinyML Anomaly Detection System for Extreme Industrial Environments. *Sensors*, 23(4), January 2023. Number: 4.
3. Hatim Bamoumen, Anas Temouden, Nabil Benamar, and Yousra Chtouki. How TinyML Can be Leveraged to Solve Environmental Problems: A Survey. In *2022 International Conference on Innovation and Intelligence for Informatics, Computing, and Technologies (3ICT)*, pages 338–343, November 2022. ISSN: 2770-7466.
4. Andhe Dharani, S. Anupama Kumar, and Peethi N. Patil. Object Detection at Edge Using TinyML Models. *SN Computer Science*, 5(1):11, November 2023.

5. Jesse Dodge, Taylor Prewitt, Remi Tachet Des Combes, Erika Odmark, Roy Schwartz, Emma Strubell, Alexandra Sasha Luccioni, Noah A. Smith, Nicole DeCario, and Will Buchanan. Measuring the Carbon Intensity of AI in Cloud Instances. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 1877–1894, Seoul Republic of Korea, June 2022. ACM.
6. Dr. Lachit Dutta and Swapna Bharali. TinyML Meets IoT: A Comprehensive Survey. *Internet of Things*, 16:100461, December 2021.
7. Zhong Fan and Ran Liu. Investigation of machine learning based network traffic classification. In *2017 International Symposium on Wireless Communication Systems (ISWCS)*, pages 1–6, August 2017. ISSN: 2154-0225.
8. Harzing, Anne-Wil. Publish or Perish, 2007.
9. Gian Marco Iodice. *TinyML Cookbook: Combine artificial intelligence and ultra-low-power embedded devices to make the world smarter*. Packt Publishing Ltd, April 2022.
10. Asharul Islam Khan and Salim Al-Habsi. Machine Learning in Computer Vision. *Procedia Computer Science*, 167:1444–1451, January 2020.
11. Pete Warden and Daniel Situnayake. *TinyML*. O'Reilly, 2019. ISBN: 9781492051992.
12. Shvetank Prakash, Matthew Stewart, Colby Banbury, Mark Mazumder, Pete Warden, Brian Plancher, and Vijay Janapa Reddi. Is TinyML Sustainable? *Communications of the ACM*, 66(11):68–77, November 2023.
13. Visal Rajapakse, Ishan Karunanayake, and Nadeem Ahmed. Intelligence at the Extreme Edge: A Survey on Reformable TinyML. *ACM Computing Surveys*, 55(13s):282:1–282:30, July 2023.
14. Aimee van Wynsberghe. Sustainable AI: AI for sustainability and the sustainability of AI. *AI and Ethics*, 1(3):213–218, August 2021.
15. Chandrasekar Vuppalapati, Anitha Ilapakurti, Sharat Kedari, Jaya Vuppalapati, Santosh Kedari, and Raja Vuppalapati. Democratization of AI, Albeit Constrained IoT Devices & Tiny ML, for Creating a Sustainable Food Future. In *2020 3rd International Conference on Information and Computer Technologies (ICICT)*, pages 525–530, March 2020.
16. Shaojie Zhuo, Hongyu Chen, Ramchalam Kinattinkara Ramakrishnan, Tommy Chen, Chen Feng, Yicheng Lin, Parker Zhang, and Liang Shen. An Empirical Study of Low Precision Quantization for TinyML, March 2022. arXiv:2203.05492 [cs].

Authors

Marco Wagner received his diploma in Automotive Systems Engineering from Heilbronn University and his PhD in Computer Science from University of Koblenz-Landau, Germany. He worked in several domains in industry both in research, development and managing positions for almost 10 years. Recently, he moved back into academics and took over the professorship for AI in technical systems at Heilbronn University, Germany. His research interests include both embedded systems design and machine learning in combination with approaches towards more sustainability.